Categorization of Government Information (CGI) Working Group
U.S. Federal Interagency Committee on Government Information

# Defining What Government Information Is To Be Categorized: Statement of Requirements

# Revised May 13, 2004

This document is located on the Internet at
http://www.gpoaccess.gov/cgiwg/pdf/cgiwgroup/revMay2004.pdf

Comments on this document may be sent to Gil Baldwin, U.S. Government Printing Office (e-mail: ebaldwin@gpo.gov).

### 1. Objective

The U.S. Federal Government seeks to enhance search interoperability by adopting common standards, as required under the E-Government Act of 2002, Section 207 "Accessibility, Usability, and Preservation of Government Information." Paragraph 207(d)(1) of the E- Government Act (44 U.S.C. Chapter 36) requires that the Interagency Committee on Government Information (ICGI) submit recommendations to the Director of the Office of Management and Budget (OMB) on:

- the adoption of standards, which are open to the maximum extent feasible, to enable the organization and categorization of Government information in a way that is searchable electronically, including by searchable identifiers; and in ways that are interoperable across agencies;
- the definition of categories of Government information which should be classified under the standards; and
- determining priorities and developing schedules for the initial implementation of the standards by agencies.

This document outlines the requirements for defining categories of U.S. Federal Government information that should be classified under the CGIWG recommended standards.

### 2. Background

#### 2.1 Overview

A work planning session was held on January 20, 2004 at the White House Conference Center in Washington D.C. Among the participants in that work planning session there was strong agreement that the scope of CGI encompasses all Federal Government

information, including tangible resources as well as intangible (online electronic) resources.

**2.2 Definitions**

Historically several relevant definitions of Government information have been codified. The narrowest of these definitions is that of "Government publication" found at 44 U.S. Code 1901, the governing statute for the Federal Depository Library Program:

> As used in this chapter "Government publication" means informational matter which is published as an individual document at Government expense, or as required by law.

This language, derived from the paper documents era, excludes the growth areas of Federal electronic information. Entire categories of Government information, such as dynamic data, audio or video files, statistical data, remote sensing data, and more are ignored by a definition that emphasizes the fixed "documentary" nature of legacy print products.

Clearly a broader definition is needed. The broadest relevant statutory definition is that for Federal records, found at 44 USC 3301:

> … "records" includes all books, papers, maps, photographs, machine readable materials, or other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of data in them. Library and museum material made or acquired and preserved solely for reference or exhibition purposes, extra copies of documents preserved only for convenience of reference, and stocks of publications and of processed documents are not included.

This language underlies the work of the National Archives and Records Administration (NARA) safeguarding the records on which the American people depend for documenting their individual rights, for ensuring the accountability and credibility of their national institutions, and for analyzing their national experience. Today more of these records are being electronically created and maintained than ever before, and NARA anticipates exponential growth in the number of electronic records to be maintained and made accessible in the coming years. However, this broad definition of categorizable information, which includes potentially billions of email messages and other work products, strains the boundaries of the ICGI Working Group's charter.

We need to identify a manageable middle ground which, while recognizing the need to protect national security interests and personal privacy rights, is sufficiently broad to

encompass information dissemination formats yet to be invented, but focuses on published information. Such language is found in the 44 USC 3502 definition of public information, at paragraph 12:

> [T]he term "public information" means any information, regardless of form or format, that an agency discloses, disseminates, or makes available to the public.

A consequence of adopting this definition could be to exclude from CGI information products that were produced for an internal agency audience, but that are also of public interest. This concept is codified in 44 USC 1902, which requires that:

> Government publications, except those determined by their issuing components to be required for official use only or for strictly administrative or operational purposes which have no public interest or educational value and publications classified for reasons of national security, shall be made available … for public information.

The following definition, which is focused on information products, of interest to the public, *produced by or for* the Government, is recommended:

> The term "categorizable Government information" means any information product, regardless of form or format, that an agency discloses, publishes, disseminates, or makes available to the public, as well as information produced for administrative or operational purposes that is of public interest or educational value. This includes information created or exchanged within or between agencies, and information that is or may be expected to be subject to FOIA requests. Not included are Federal government information holdings explicitly provided in law as so constrained in access that even a reference to the holding is kept from public view for a specified period of time.

However, it must be recognized that cases will occur in which the publishing agency may limit access to the descriptive metadata about certain products to certain audiences for a specified period of time, due to security, privacy, or other records management reasons..

The goal of agreeing upon, and ultimately implementing, a definition of what information is to be categorized, is to enable users to obtain a predictable body of search results, of similar granularity across varying communities of interest.

## 3. Assumptions and Constraints

### 3.1 Scope of Definition

Searchers of government information need to find tangible resources (i.e. printed documents, maps, CDs, or DVDs) as well as intangible (online electronic) resources produced by or for the Government. The definition of resources to which CGI is applicable should not be so all-encompassing as to be unmanageable. For that reason it is recommended that information products *about* the Government, such as television news

coverage of Government activities, be excluded. For similar reasons, applying CGI to objects *owned by or loaned to* the Government, such as museum artifacts, should be excluded. An overly broad definition of CGI-eligible resources risks creating a requirement so burdensome to the Government that the goal of improved public access will be jeopardized.

### 3.2 Limited Exclusion for Restricted Information Resources

The Federal government generally does not constrain access to or use of its holdings and the data and information are regarded as being in the public domain. Yet, there are a range of constraints that may apply to any particular holding. Use constraints such as copyright restrictions may apply in certain cases specifically allowed under law, such as patents. Access constraints may apply to certain security classified information, proprietary information, personal information, litigation-related information, and other particular cases. For example, there is certain information for which access is restricted to authorized public citizens such as (1) Information restricted to private citizens eligible to receive that data, (2) information limited to government contractors, (3) information limited to state and local governments. It is important that these types of information also be included under the scope of CGIWG categorization.

While all government information will be not accessible to the public, but awareness of its existence and the restrictions on such access should be. Even when information may be withheld from disclosure, publication, or dissemination the public has a right to know about its existence. The only information out of scope for this discussion are those few Federal government information holdings explicitly provided in law as so constrained in access that even a reference to the holding is kept from public view.

### 3.3 Distributed Information; Centralized Services

There was strong agreement that government information will always be highly distributed among many resources that are separately maintained. The CGI challenge is to define common standards that support interoperability. CGI should enlighten government efforts both to organize information for its own uses and to address the public need for cross-agency coherence in describing "how to get it".

### 3.4 Implementation Burden and Costs

The U.S. Federal Government expends massive resources on collecting, compiling, disseminating, and preserving government data and information. Support of a search service standard would entail additional cost initially, but that addition should be a small percentage of the overall cost. In establishing Government-wide standards for interoperability and other requirements, consideration must be given to burdens, including both costs and operational difficulties, that will flow from the standards. Operational difficulties arise when the organization posting information is ill-suited to implement a standard, or when individual information products do not conform to the expected structure or format.

There is an ongoing operational cost to government in supporting any search service standard. For every major type of information resource offered through the standard search, someone familiar with the holdings must identify what equivalences exist between the standard search interface and the locally held information. This one-time "semantic mapping" task is typically handled by a system administrator, and is essentially the same function as required for setting up any non-standard search interface.

## 4. Major Stakeholders

**[Note: This section adopted from the draft "Statement of Requirements for Search Interoperability."]**

### 4.1 The Public and Non-Government Organizations

The right for the public to have direct access to data and information resources held by governments is a long-standing tenet of public policy and is codified in laws such as the Freedom of Information Act and the Privacy Act, among many others. Governments have a responsibility to facilitate such direct access to the array of mechanisms for data and information access now being maintained. Agreement on a reasonable definition of what Government information is to be categorized will advance meeting this responsibility in a substantial manner.

The various public communities served by governments can obtain information either directly from government organizations or through intermediaries. Traditionally, libraries and information services have played a major intermediary role in public access to government information, and non-government organizations (NGO's) have a similar stake in government information access mechanisms. Public access to government information is increasingly being accomplished through public network facilities such as Web pages and Internet search engines. Such access is typically regarded as "dis-intermediated", although of course a degree of intermediation is inherent in the choice of technologies and information selection criteria.

### 4.2 Libraries and Information Service Providers

Although many people perform casual searching on their own, the intermediation roles of libraries, information services, and NGO's will continue into the indefinite future. Much of the nation's public continues to rely on trained searchers and librarians to provide essential services in access to government information. Public access is supported by specialized training in library schools and by a massive and pervasive infrastructure. For instance, the United States has more than 120,000 libraries, including over 1,250 Federal depository libraries.

Online information services (e.g., Lexis/Nexis, Chemical Abstracts Service, Dow Jones News Retrieval) represent another major community of practice. These services typically

provide fee-for-service search access and for obvious commercial reasons they have less incentive than libraries to support open search standards. Yet, online information services are often major intermediaries for government holdings and most offer support for the international standard search service adopted by libraries.

### 4.3 Search Technology Vendors

The development of Internet search engines can be traced to the advent of Web crawling technology. Because Web pages were constructed using HyperText Markup Language (HTML) and contained a high proportion of unstructured, document-like information, content was mostly indexed for search using full-text search technologies. For some years, debates raged over the idea that full-text search engines offered an unbeatable price/performance ratio in comparison to more traditional cataloging techniques. Today, most Internet search technologies offer a combination of wholly automated and machine-aided cataloging techniques, and treat Web content as semi-structured information. This responds to the user requirement for good "precision" as well as "recall". This extension of Internet search technologies to handle structured information is driven commercially by the requirement for search interoperability within Intranets, i.e., company internal databases, directories, etc.

### 4.4 Government Organizations

Every government organization holds a wide variety of data and information resources and maintains a wide range of directories and other data and information locators. Data and information may be in the form of paper or electronic documents, budget tables, e-mail files, audio and video files, databases, and data systems of all kinds. U.S. Federal agencies and many state agencies are required to maintain locators for information and collections of information, which may be or may not be publicly available. This mandate is rooted in public policy interests for government transparency, accountability, and protection of privacy.

These resources are composed of discrete data elements and such data is only useful if information about the data element is also readily available. This kind of documentary information (sometimes called "metadata") is typically held in yet another resource known as a "data dictionary", "metadata catalog", "repository", or "registry". Making the CGI concept function will require U.S. Government organizations to apply this metadata to the information products within their purview, preferably before or at the moment of publication.

## 5. Process for Identifying Requirements

[Section not applicable to this topic]

## 6. Major Requirements

In order for the CGI effort to add value for the information user, it should meet several general major requirements. Most importantly, it should enhance public access to Government information resources. The users should be able to expect a predictable level of granularity among the search returns from decentralized data sources. The CGI initiative must be a realistic mandate for Government entities, many of which operate with less than optimal levels of funding or IT support, to carry out. And CGI must be compatible with the installed base of existing information characterization and retrieval mechanisms, many of which represent an enormous investment of public funds.

### 6.1 Enhances Public Access

Public access to Government information can be enhanced by a CGI initiative if there is broad buy-in and implementation among the agencies. Even with the scope limits recommended in this paper, the potential universe of in-scope resources is vast and rapidly expanding. For the public to use, and ultimately support through tax dollars and/or user fees, the CGI applications there must be demonstrable results that make using it worth their while. For the public, the greatest demonstrable results will accrue from applying CGI processes to publicly available information. Agencies may benefit from an improved ability to identify and locate intragovernmental or other non-public or restricted information resources.

### 6.2 Equivalent Level of Granularity

User acceptance of the CGI application will be enhanced if the information products identified represent an equivalent level of granularity from one provider to another. A CGI system will not gain broad support and acceptance if some providers are characterizing information at the journal article level while others are representing entire data bases collectively. The scope of what products are categorized, and the level of resource categorized, will likely need to be mandated in order to achieve a broad based implementation.

### 6.3 Executable Mandate

A critical mass of CGI-processed information resources is necessary to reach the tipping point to wide usage and general commitment to the initiative. For agencies to support what is likely to an unfunded information processing mandate, the benefits of the mandate need to be obvious and its execution simple. In the interest of simplicity of understanding, and to accrue of critical mass of CGI-processed resources in a reasonable time, the boundary of what is in scope for categorization needs to be defined carefully. A recommendation that CGI processes and metadata control be applied to all information resources or related artifacts that are produced for, published by, controlled by, owned by, loaned to, or about the Federal Government the mandate will collapse of its own weight.

### 6.4 Compatibility with Existing Mechanisms

Defining what public information resources are in scope for CGI applications must take into account the existing and developing mechanisms for access. These include:

- Direct Access to Data
- Library Catalog Search
- Internet Search Engines
- Government Locators
- Search Request/Response services
- Semantic Mapping

These mechanisms are more fully discussed in the "Statement of Requirements for Search Interoperability."

## 7. Notes and References


[FEA DRM] Office of Management and Budget, Federal Enterprise Architecture: Data and Information Reference Model, [not yet published], March 2004

[ISO] International Organization for Standardization, ISO/TR 15489-2:2001, Information and documentation -- Records management -- Part 2: Guidelines, available at: http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=35845&ICS1=1&ICS2=140&ICS3=20

[OAI] Open Archives Initiative, http://www.openarchives.org/ , 2003.

[OASIS] The OASIS E-Government Technical Committee recommendation concerning Search Service interoperability is available at http://www.oasis-open.org/committees/download.php/4274/wd-egov-searchservice-01.pdf

[USC] United States Code, 2000 edition. Title 44, Public Printing and Documents, Chapter 19, Section 1901, Definition of Government publication, available at http://www.gpoaccess.gov/uscode/index.html

[USC] United States Code, 2000 edition. Title 44, Public Printing and Documents, Chapter 19, Section 1902, Availability of Government publications … , available at http://www.gpoaccess.gov/uscode/index.html

[USC] United States Code, 2000 edition. Title 44, Public Printing and Documents, Chapter 19, Section 3301, Definition of records, available at http://www.gpoaccess.gov/uscode/index.html

[USC] United States Code, 2000 edition. Title 44, Public Printing and Documents, Chapter 19, Section 3501, Definition [of public information], available at http://www.gpoaccess.gov/uscode/index.html

[WSIS-1] The Plan of Action of the World Summit on the Information Society (WSIS) commits governments to: "a) Implement e-government strategies focusing on applications aimed at innovating and promoting transparency in public administrations and democratic processes, improving efficiency and strengthening relations with citizens; b) Develop national e-government initiatives and services, at all levels, adapted to the needs of citizens and business, to achieve a more efficient allocation of resources and public goods; and c) Support international cooperation initiatives in the field of e-government, in order to enhance transparency, accountability and efficiency at all levels of government."

[WSIS-2] The Plan of Action of the World Summit on the Information Society (WSIS) asserts: "Standardization is one of the essential building blocks of the Information Society. There should be particular emphasis on the development and adoption of international standards... International standards aim to create an environment where consumers can access services worldwide regardless of underlying technology."